DE GRUYTER OLDENBOURG Journal of Economics and Statistics 2017; 237(5): 457–467

Data Observer **Open Access**

Michael Weinhardt*, Alexia Meyermann, Stefan Liebig and Jürgen Schupp

The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Content, Design and Research Potential

DOI 10.1515/jbnst-2015-1044

1 Introduction

The data set presented in this article results from a project to produce a Linked-Employer-Employee data set for the Socio-Economic Panel (SOEP).[1] In 2012/13, a survey of German employers was conducted using face-to-face and paper-and-pencil interviews (N = 1,708; response rate = 30.1%). Establishments were sampled based on address information provided by employed participants from the SOEP. The SOEP is a longitudinal study of German households that are representative of the German population, repeatedly surveying about 20,000 individuals and 10,000 households each year.[2] The information obtained from both surveys can be linked in order to create a linked employer–employee data set concerning organizational context and individual outcomes (N = 1,834, mostly one employee per employer). The information collected in the LEE

1 Designated "SOEP-LEE" (The Linked Employer–Employee Study of the Socio-Economic Panel), the study involved cooperation between the SOEP department at DIW Berlin and Bielefeld University. The project ran from January 1, 2012, to December 31, 2013, and received funding from the Wissenschaftsgemeinschaft Leibniz e. V. for these two years (SAW 2012-SOEP-2).
2 www.diw.de/soep

***Corresponding author: Michael Weinhardt,** Fakultät für Soziologie, Universität Bielefeld, Postfach 100 131, 33501 Bielefeld, Germany, E-mail: michael.weinhardt@uni-bielefeld.de
Alexia Meyermann, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstraße 29, 60486 Frankfurt am Main, Germany, E-mail: meyermann@dipf.de
Stefan Liebig, Fakultät für Soziologie, Universität Bielefeld, Postfach 100 131, 33501 Bielefeld, Germany, E-mail: stefan.liebig@uni-bielefeld.de
Jürgen Schupp, German Institute for Economic Research (DIW), Mohrenstr. 58, 10117 Berlin, Germany; and Freie Universität Berlin, Kaiserswerther Str. 16-18, 14195 Berlin, Germany, E-mail: jschupp@diw.de

study reported enrich and enhance the existing individual-level and household-level SOEP data with supplemental data about the workplace and the employees' working conditions. In contrast to the SOEP core study, the SOEP-LEE data set contains more detailed and independent information concerning the work context. This way, the LEE data can be used to investigate the organizational impact on the genesis of social inequalities and the individual development of the life course. The SOEP-LEE study specifically sought to obtain information about inter-organizational as well as intra-organizational heterogeneities such as forms of employment (part-time, full-time), temporary work, and similar atypical forms of employment, as well as about other factors, such as gender composition, the age of the employees, and the wage structure of the establishment. The overall aim was to investigate social inequalities and their relation to employers and organizations (e. g., to determine how organizational structures and practices influence social inequality at the individual level). A detailed project report of the study can be found in Weinhardt et al. (2016).

2 Topics covered in the SOEP-LEE study

The SOEP-LEE questionnaire was designed to measure the role organizations play as both contexts and actors in the generation of social inequality, taking into account additional information available from the SOEP on the individual level. Questions cover (1) job-specific practices and structures concerning central dimensions of inequality such as income, promotion prospects, and gratifications, (2) inter-organizational variance, assessed by measuring aggregated characteristics on the organizational level, as well as (3) intra-organizational variance (e. g., differences between groups of employees within the establishment). Using this information, questions of social inequality may be addressed, first, *within* these establishments by looking at intra-organizational differences in income/wages, job mobility (career opportunities), and working conditions (e. g., working hours, climate). Second, inequalities may be investigated *between* establishments by comparing organizations' personnel policies, strategies, demographics, and the overall economic and financial situation of the establishment. Thus, researchers might want to focus on the strategies, practices, and structures that are in place within establishments or between establishments, investigating to what extent employees belong to privileged groups and how groups differ along different dimensions of inequality.

Because the questionnaire was also constructed to match the information collected in the 2011 wave of the SOEP survey, the SOEP-LEE data set

substantively enhances the information on individual work contexts and working conditions of respondents to the SOEP survey. The SOEP itself offers rich information on a wide range of inequality dimensions, such as education, health and income. Social and economic characteristics (personal finances) are included, as well as measures of attitudes and personality traits. The SOEP-LEE data set therefore offers new research opportunities in the field of organizational inequality and beyond.

The establishment questionnaire consisted of a fully structured paper-and-pencil questionnaire (or paper-and-pencil interviewing [PAPI]), with interviewers administering the questionnaire face-to-face and the option of self-completion, if requested by the respondent. This information can be analyzed in conjunction with the establishment data after the two data sources have been linked (see Section 5). The resulting questionnaire comprised 61 questions (161 items) and took an average of 40 minutes to complete. For more detailed information on the topics addressed by the questionnaire and the constructs underlying specific operationalizations, see Table 1. The full version of the original questionnaire, as well as an English translation, can be found in Weinhardt (2016b).

3 The SOEP-LEE data as a representative sample of German employers

The SOEP-LEE study differs from many other surveys in the way its sample was derived. Although it can be thought of as a probability sample of German establishments, the sample was not drawn from a sampling frame of establishments; rather, a random sample of employees coming from the SOEP was used as the basis for the study, employing the so called "employee first method" (Kmec 2003). In the SOEP 2012, SOEP respondents who had reported to be employed ("*abhängig beschäftigt*") in 2011 were asked for the name and contact details of their employers.[3] The resulting list of addresses was used as the gross sample for the subsequent, separate establishment survey of employers. An

3 This method has been successfully pretested in 2007 (Meyermann et al. 2009). Under the condition that the employee sample, the first sampling stage, is fully representative, each establishment in the population has a non-zero probability of being selected which becomes known only by after the sampling stage by collecting information on the number of employees in the establishment. However, proxies on establishment size are usually available in every employee survey.

Table 1: Topics and corresponding questions in the SOEP-LEE establishment questionnaire.

Type of organization; basic characteristics	Question number(s)
Type	Q1
Public vs. private	Q1, Q6
Size	Q2, Q21, Q21a, Q26
Age	Q5
Industry	Q6
Owner	Q17, Q20
Legal form	Q18, Q18a
Nonprofit vs. for-profit	Q19
Economic and financial situation of the establishment	
General questions	Q10–Q12
Problems and challenges	Q15, Q16
Financial status (turnover, share of staff costs)	Q21–Q23
Staff/human resources (HR) policy	
Flexibilization strategy	Q13, Q14
HR problems	Q24
Detailed staff structure	Q25–Q29
Openings and progression	Q33–Q35
Questions on wages/income	
Questions measuring various forms of remuneration, including collective agreements	Q38–Q46
Direct measures of inequality *within* the organization	Q31, Q32, Q46, Q52–Q54
Single theoretical constructs	
Firm-internal labor market	Q37
Organizational culture and climate	Q47 (12 items)
Differentiation, centralization	Q4, Q7–Q9, Q36
Transparency	Q40, Q48
Formalization	Q39, Q49, Q50
Participation (employee representation)	Q51, partly Q47
Export orientation	Q11
Autonomy of a single establishment/belonging to a larger organization	Q2, Q4
Organizational slack	Q12
Working hours and work–life balance	Q52–Q54
Other	
Informed consent on record linkage, report offering	Q3, Q30, Q62
Respondent characteristics	Q55–Q61

employer address was fielded for 5,919 employees, or 52.7 percent of the original 11,229 individuals who were employed in 2011. For these 5,919 establishments, 1,708 interviews could be achieved, resulting in an adjusted response rate of 30.1 percent at the organizational level. The information collected from the

employer survey can be matched with the survey data on the individual employees to form a linked employer–employee (LEE) data set for 1,834 employees.

In this kind of sampling design, respondents drop out at several points during the data collection process for a wide range of reasons between the original person-level sample of employees in 2011 and the fielded sample for the employer survey in 2012 (Bechmann/Sleik 2016). Such reasons include panel attrition, failure to provide an employer address, inaccuracy of addresses provided or unwillingness of establishments to take part in the survey. For example, at the employee level, in 543 cases, the address questionnaire was not returned; in 1,182 cases (12.8 %), the address was incomplete or missing (possibly refused); and an additional 125 addresses could not be verified and were therefore dropped. At the employer level, 505 cases were excluded as duplicates (i. e., more than one SOEP respondent reported working for this employer), but obviously the employer would only be interviewed once. Due to data protection issues a batch of 949 addresses had to be dropped because the establishments in question were reported to have fewer than five employees[4]; self-employed persons were excluded as well.

It becomes apparent that there is substantial cumulative drop-out over the whole process of the data collection process. Therefore, problems of sample selectivity may arise which are considered here briefly. When using the employee-first approach, it is important to start from a sample of individuals that represents the population or, more precisely, the workforce within the population. The SOEP can be considered a suitable and valid starting point for the employee-first method owing to its overall representativeness. To further investigate nonresponse at the following drop-out stages in the sampling design, we analyzed unit nonresponses at two levels: (1) employees not providing the names and addresses of their employers and (2) establishments not participating in the establishment survey. The analyses showed that the response process and the losses that occurred at these two steps of the sampling process were not entirely random but exhibited patterns of systematic nonresponse. This is important information and must be kept in mind when analyzing the data. It is also interesting to see how the different effects at the two levels interacted with each other: for example, the larger an organization, the more likely it was that the SOEP respondents would provide this contact information, but the less likely it was that the establishment would actually respond to the survey request. Here, the two effects at the two levels potentially canceled each other out. However, cumulative effects were also present: public sector employees

4 This restriction was established to lower the potential risk of re-identification of establishments.

were more likely to name their employer, and public sector establishments were more likely to respond, potentially leading to overrepresentation of the public sector (especially educational agencies) in the SOEP-LEE sample. Thus, analysts of the data should look into the sample composition and the factors influencing responses before they run their analyses and interpret their results.

Despite these concerns, the resulting employer sample is basically a cross-section of German establishments, including employers throughout the country, across all lines of businesses, from the private, public, and tertiary sectors (associations and foundations). Table 2 shows the distribution of establishments that took part in the study according to number of employees, industry (WZ 2008 classification), and region (federal states). For comparison, this table also contains official data from German employment statistics (Federal Employment Agency [FEA]). Before weighting, numbers reflect the distribution at the employee level, after weighting the distribution at the establishment level. Looking at the size distributions, we can see that as a result of the sample design, the chance of larger establishments being selected is much greater relative to their share in the actual population. In order to generalize the population of establishments, the data must be weighted. The design weight used here consists of the inverse of the establishment size to account for unequal selection probabilities of the establishments. After weighting, the percentages of the SOEP-LEE sample come close to the official data, which may serve as the actual percentages in the population of establishments. This supports the validity of the sampling procedure when unequal selection probabilities are taken into account. Still, differences between the SOEP-LEE sample and the official statistics remain and should be kept in mind when interpreting results (see Table 2).

4 Data quality, editing and anonymization

Data quality and measurement error were investigated through the interviewers' evaluations of the interview situation and an analysis of item nonresponse in the final data set. The interviewers noted only a few problems with the response process and the interview situation. Overwhelmingly, they perceived the response persons to be knowledgeable and accurate, and according to the interviewers, even complex and burdensome questions were answered to the best of their abilities. Overall, the analysis of item nonresponses showed that missingness in the data is low. The item nonresponse rate was high (up to 30 %) for only a few items, especially those concerning financial information about the establishment. Yet, all in all, missing information at the item level was not a

Table 2: Comparison of SOEP-LEE sample and official statistics.

Number of employees	Employee level				Establishment level*		
	SOEP-LEE		FEA		SOEP-LEE	FEA	
	N	%	%	Diff.	%	%	Diff.
6–9	121	7.2	7.1	0.1	31.7	36.9	−5.2
10–19	234	13.9	10.7	3.2	33.6	29.9	3.7
20–49	335	19.9	15.7	4.2	21.7	19.4	2.3
50–99	269	16	13.4	2.6	7.6	7.3	0.3
100–199	261	15.5	13.7	1.8	3.7	3.7	0
200–249	63	3.7	4	−0.3	0.5	0.7	−0.2
250–499	142	8.4	11.6	−3.2	0.8	1.3	−0.5
≥ 500	258	15.3	23.7	−8.2	0.5	0.7	−0.2
Total	1,683	100	100	23.6*	100	100	12.4*
Mean (absolute differences)	210			3			1.6
Industry (WZ 2008)	**N**	**%**	**%**	**Diff.**	**%**	**%**	**Diff.**
Agriculture, forestry, and fishing	34	2.1	0.8	1.3	3.6	2.6	1
Mining, quarrying; electricity; water supply, sewerage	38	2.4	1.9	0.5	1.4	0.8	0.6
Manufacturing	332	20.4	22.5	−2.1	16.9	9	7.9
Construction	72	4.4	5.8	−1.4	8.9	10.7	−1.8
Wholesale, retail trade; repair of vehicles	151	9.3	14.4	−5.1	14.2	20.3	−6.1
Transporting and storage	48	2.9	5.1	−2.2	2.3	4	−1.7
Accommodations and food service activities	46	2.8	3.1	−0.3	4.6	7	−2.4
Information and communication	36	2.2	3	−0.8	2.9	2.6	0.3
Financial and insurance activities	39	2.4	3.5	−1.1	1.1	3	1.9
Real estate activities; professional, scientific, and technical activities	95	5.8	13.6	−7.8	5.6	17	−11.4
Public administration and defense; compulsory social security	199	12.2	6	6.2	4.8	1.5	3.3
Education	199	12.2	3.8	8.4	12.7	2.6	10.1
Human health and social work activities	287	17.6	12.6	5	17.3	10.3	7
Other service activities; arts, entertainment and recreation	55	3.4	3.8	−0.4	3.4	8.7	−5.3
Total	1,631	100	100	42.6*	100	100	60.8*
Mean (absolute differences)				3			4.3>
Federal state	**N**	**%**	**%**	**Diff.**	**%**	**%**	**Diff.**
Schleswig-Holstein	65	3.8	3	0.8	3.9	3.6	0.3
Hamburg	21	1.2	2.9	−1.7	5	2.4	2.6

(*continued*)

Table 2: (*continued*)

Number of employees	Employee level				Establishment level*		
	SOEP-LEE		FEA		SOEP-LEE	FEA	
	N	%	%	Diff.	%	%	Diff.
Lower Saxony	141	8.3	8.9	−0.6	10.3	9.1	1.2
Bremen	18	1.1	1	0.1	1.2	0.7	0.5
North Rhine-Westphalia	273	16.1	21	−4.9	12.8	19.9	−7.1
Hesse	137	8.1	7.9	0.2	6.8	7.5	−0.7
Rhineland-Palatinate	80	4.7	4.4	0.3	5.8	4.9	0.9
Baden-Württemberg	232	13.7	14	−0.3	13.3	13.1	0.2
Bavaria	267	15.7	16.6	−0.9	16	16.8	−0.8
Saarland	17	1	1.3	−0.3	0.8	1.2	−0.4
Berlin	41	2.4	4.1	−1.7	2.4	4.1	−1.7
Brandenburg	80	4.7	2.7	2	5.4	3.1	2.3
Mecklenburg-Vorpommern	43	2.5	1.9	0.6	2.8	2.3	0.5
Saxony	118	6.9	5.1	1.8	8	5.5	2.5
Saxony-Anhalt	82	4.8	2.7	2.1	4.8	2.8	2
Thuringia	85	5	2.6	2.4	4.5	2.9	1.6
Total	1,700	100	100	20.7*	100	100	25.2*
Mean (absolute differences)				1.3			1.6

Source: SOEP-LEE Study and Federal Employment Agency (FEA); year of reference: 2011; establishments with more than five employees.
Notes: The SOEP-LEE sample excludes establishments with fewer than five employees, but the official statistics (FEA) include all establishments with one to five employees. Therefore, the first size category was excluded for computation of percentages. *Here, SOEP-LEE data is weighted to take into account the unequal selection probability based on the size of the establishments and to allow valid comparisons at the establishment level.

major problem in the SOEP-LEE study. Both the inspection of item nonresponse rates and the interviewer observations concerning the response process appear to confirm that the quality of the data overall was good.

In order to achieve high data quality, the data underwent extensive checking and editing procedures. Many of the establishments were contacted by telephone to verify certain data and to control the work of the interviewers. Little of the information had to be revised at this step. In order to ensure data privacy, the interviews were anonymized by the survey agency by separating address data from survey data. Further anonymization was done by the SOEP-LEE team with advice from experts of the Research Data Center for Business and Organizational Data (RDC-BO) at Bielefeld University. Several items were identified that risked the re-identification of the participating establishments if they

were not edited in order to ensure privacy. The overall goal of anonymization was to ensure safe long-term access to and use of the data by third parties (data sharing) and the scientific community. This should be considered when analyzing the data set. The questionnaire/code plan (TNS Infratest Sozialforschung 2016) and the data manual (Weinhardt 2016a) also indicate anonymized variables.

5 Data dissemination and linkage

Linking survey data to establishment data is possible for 1,834 individuals (110 establishments with more than one SOEP employee; the maximum number of employees per establishment is six). For substantial analyses, it is important for researchers to be clear about the level of their interest: the individual or the establishment. All substantial analyses (e. g., regression analyses) should probably include the size of the establishment as a covariate because (a) it is a major determinant of many processes at the establishment level and can potentially be correlated with a wide range of variables, and (b) the probability of selection depends heavily on establishment size. In computing standard errors, one has to account for the fact that, for a fraction of the combined sample at least, more than one SOEP respondent is nested within one establishment.

The SOEP-LEE establishment data set (Liebig/Schupp 2014, DOI:10.7478/s0549.1.v1) is available for secondary use at two data archives in Germany, the SOEP-RDC at DIW Berlin and the RDC-BO at Bielefeld University. The dissemination of these data, along with the normal SOEP data, is restricted owing to the sensitivity of the data and the risk that individual establishments might be identified. Researchers can analyze the entire database either during a research stay at the SOEP or at the RDC-BO. All outputs will be checked to ensure that the data provided remain confidential. To facilitate research and analysis, the questionnaire and code plan is provided (Weinhardt 2016b) together with a data manual describing the establishment data set, including frequency distributions of the variables (Weinhardt 2016a).

To enrich the study by adding information drawn from administrative data, the SOEP-LEE data on employers were linked to data from the Establishment History Panel (BHP) of the Institute for Employment Research (IAB Nuremberg). The BHP consists of aggregated data on employees who are subject to social insurance contributions and on their incomes as reported by their employers to the Federal Employment Agency (Bundesagentur für Arbeit [BA]). Aggregated to the establishment level, these data contain information about the income, sex, and education composition of the establishment and

thus could expand the data in the SOEP-LEE study. Of the 1,708 establishments interviewed in the SOEP-LEE study, 587 establishments (35.2 %) gave their consent to allow the records to be linked to the administrative data held by the IAB. Of these, 443 establishments (75.5 %) could eventually be linked to IAB establishment data (Eberle/Weinhardt 2016).[5]

6 Publications and presentations

In order to distribute these data to researchers and the scientific community, we have prepared posters and presentations specifically tailored for different occasions, scientific workshops, and conferences. The SOEP-LEE data have also been used for theses written for bachelor's, master's, and doctoral degrees (a full list is available in Weinhardt et al. 2016). In addition, the data is used for the DFG funded research project: "Die ambivalente Bedeutung betrieblicher Strukturen für die Erklärung sozialer Ungleichheit zwischen Frauen und Männern – Analysen mit dem SOEP-LEE" (financed by the German Research Foundation (DFG) for 2016–2018, principal investigator: Prof. Dr. Anne Busch-Heizmann, Universität Dusiburg-Essen).

7 Summary

The SOEP-LEE study, employing the "employee-first" method was implemented based on a large sample of employees, the SOEP, and resulted in a relatively large sample of employers, covering a variety of organizations that were subsequently interviewed. The resulting data provides a variety of new options for analysis: (1) Social inequality measures found in the SOEP individual data can be analyzed by asking how they are connected to workplace characteristics; (2) not only does the workplace data set itself provide measures of inequality, but it also allows us to distinguish between inter-establishment and intra-establishment inequality measures. Until now, no data were available that link the data from household-level and individual-level longitudinal surveys to survey data concerning the role of the workplace in life outcomes. This data set will allow us to examine these three levels simultaneously. In a wider perspective, the SOEP-

5 So far, identifiers alone have been matched to facilitate linkage. In order to use the actual linked data of both data sources, one must apply separately to the IAB and comply with strict data confidentiality regulations regarding BHP data.

LEE data is unique within Germany; no linked data set of the size and richness of the information contained here has so far been available. It will substantially augment the information available through the SOEP regarding the work contexts and working conditions of the SOEP respondents. Thus, the SOEP-LEE data set opens up new possibilities for a wide range of secondary analyses to answer innovative research questions from the fields of economics and social sciences, such as organizational determinants of inequalities in health and income or their impact on work-family conflicts.

Acknowledgments: We would like to thank everyone involved in the project for their efforts and support, from the students working with us to the external experts providing their advice.

Funding: The SOEP-LEE study was awarded a research grant from the Leibniz Association based on a successful proposal to the Leibniz Competition (SAW 2012-SOEP-2).

References

Bechmann, S., K. Sleik (2016), SOEP-LEE Betriebsbefragung – Methodenbericht der Betriebsbefragung des Sozio-oekonomischen Panels. SOEP Survey Papers 305: Series B. Berlin: DIW Berlin/SOEP.

Eberle, J., M. Weinhardt (2016), Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP). German RLC Working Paper No. wp-grlc–2016–01.

Kmec, J.A. (2003), Collecting and Using Employer–Worker Matched Data. Sociological Focus 36 (1): 81–95.

Liebig, S., J. Schupp (2014), SOEP-LEE Betriebsbefragung – Die Betriebsbefragung des Sozio-oekonomischen Panels. doi:10.7478/s0549.1.v1.

Meyermann, A., J. Elsner, J. Schupp, S. Liebig (2009), Pilotstudie einer surveybasierten Verknüpfung von Personen- und Betriebsdaten. SOEP papers 170. Berlin: DIW Berlin.

Weinhardt, M., A. Meyermann, S. Liebig, J. Schupp (2016), The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Project Report. SOEP papers 829. Berlin: DIW Berlin/SOEP.

Weinhardt, M. (2016a), SOEP-LEE Betriebsbefragung – Datenhandbuch der Betriebsbefragung des Sozio-oekonomischen Panels. SOEP Survey Papers 306: Series D. Berlin: DIW Berlin/SOEP.

Weinhardt, M. (2016b), SOEP-LEE Betriebsbefragung – Erhebungsinstrumente und Datenkodierung der Betriebsbefragung des Sozio-oekonomischen Panels. SOEP Survey Papers 304: Series A. Berlin: DIW Berlin/SOEP.